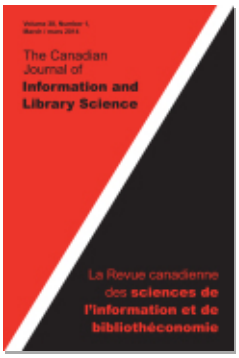
Expérimentations et évaluations en fouille de textes : un panorama des campagnes DEFT édit. by Cyril Grouin et Dominic Forest (review)



Luc Grivel

Canadian Journal of Information and Library Science, Volume 38, Number 1, March/mars 2014, pp. R3-R6 (Review)

Published by University of Toronto Press
DOI: *https://doi.org/10.1353/ils.2014.0004*

named sites, and the index entry for Boolean logic is a series of page numbers that must be checked to see what they are (and in doing this just for this review I had to force myself back to writing instead of delving into the Web). I will value both books for course preparation as there is unique content in each that will contribute, for example, to content on searching grey literature to inform health policy. However, if I could keep only one, for the quantity of new-to-me content, it would have to be *Expert Internet Searching*.

Among the content that I appreciated most for course preparation is a comparison of clustering search engines such as Carrot[2] in chapter 5 and Visual Searching in chapter 8. Chapter 11, "Academic and Other Specialized Search Engines," sets these specialized "niche market" products apart in a way that reinforces for me that information and resources must be "fit for purpose" and begs the question why we health librarians pay so much to make resources designed for academics available to health services workers without knowing for sure that these workers have enough time or capacity to search for and then synthesize multiple primary research articles to meet their day-to-day practice-based information needs.

Chapter 14, "Hints and Tips on Better Searching with Sample Search Examples," includes pointers for testing search skills, getting the most out of browsers, and setting up personal home pages for expert searchers.

I have been digging into the evolving concept of expert searching and have yet to identify recent work that provides an explicit solid, modern foundation to support expert searchers who need both a common language and standard methods to select between and then guide different search processes. This book does not set out to meet these needs, but throughout and specifically in chapter 16, "The Future of Search," it does expand the horizons of what constitutes expert searching beyond what was imaginable just a few years ago.

Jacqueline MacDonald, adjunct professor, School of Information Management, Dalhousie University

Cyril Grouin et Dominic Forest (édit.). *Expérimentations et évaluations en fouille de textes : un panorama des campagnes DEFT*. Paris, Hermès-Lavoisier, 2012. 248 pages. ISBN : 978-2-7462-3836-7. 59 €.

Cet ouvrage de la collection Systèmes d'information et organisations documentaires chez Lavoisier, coordonné par Cyril Grouin (LIMSI-CNRS) et Dominic Forest (École de bibliothéconomie et des sciences de l'information de l'Université de Montréal), rassemble dix contributions de participants aux campagnes d'évaluations DEFT (Défi Fouille de Textes) de 2006 à 2011. Ces campagnes sont organisées depuis 2007 par le LIMSI-CNRS, puis en co-organisation avec l'Université de Montréal depuis 2010.

Inspirée de la campagne d'évaluation internationale TREC, c'est la seule campagne d'évaluation francophone de fouille de textes. Elle confronte les

systèmes développés par plusieurs équipes sur un même jeu de données et en un temps limité, avec un même objectif principal : l'extraction et l'organisation automatique des informations présentes dans les textes. Cet ouvrage témoigne à la fois de la variété des problématiques et de l'amélioration progressive des méthodes et techniques de fouille de textes au fil des campagnes.

Toujours basées sur un état de l'art des méthodes selon leurs performances par rapport à une tâche donnée, les différentes contributions présentent les techniques retenues et leur stratégie, décrivent et commentent les résultats obtenus selon le processus d'évaluation de DEFT.

Si le Québec et la Belgique ont chacun une équipe parmi les contributeurs sélectionnés, on retrouve les places fortes traditionnelles des laboratoires de recherche en ingénierie linguistique française, avec le LIMSI-CNRS, le LIA, le LINA, le LORIA, l'IRISA. On peut regretter qu'il n'y ait eu qu'un seul contributeur issu de la recherche industrielle avec le CELI, devenu Ho2S, mais l'ouvrage ne décrit pas de manière précise les critères de sélection des contributeurs, si ce n'est « l'originalité des approches suivies » ou « la qualité de leurs résultats ». Ces contributions sont représentatives de l'état des techniques sur trois thématiques principales qui constituent autant de parties du livre. Chaque thématique correspond à une ou plusieurs tâches de fouille de texte comportant des difficultés particulières (existence d'un verrou technologique).

L'ouvrage comporte onze chapitres répartis en quatre parties : introduction (panorama des campagnes DEFT), les campagnes en genres et thèmes, les campagnes en fouille d'opinion, les campagnes diachroniques.

Les campagnes en genres et thèmes

Cette partie rassemble deux contributions de la campagne 2006 (identification des ruptures thématiques dans trois corpus aux propriétés structurelles très différentes (discours politiques de trois présidents de la république française, articles de lois de l'union européenne et un ouvrage scientifique)), et une contribution de la campagne 2008 (classification automatique de corpus en genres (*Le Monde* vs. Wikipédia) et en catégories (International, Politique française, Société, Économie, Sciences, Art, Littérature, Sports, Télévision)).

La première contribution, rédigée par Alain Lelu (LASEDI Université de Franche Comté) et Martine Cadot (LORIA), s'est intéressée à la détection des ruptures thématiques dans les discours présidentiels. Cette tâche pose une difficulté particulière : le nombre et la nature des catégories thématiques ne sont pas connus, y compris dans le corpus d'apprentissage. D'où l'intérêt de rechercher une synergie entre l'emploi d'une technique de classification non supervisée avec une technique d'apprentissage supervisé. La contribution suivante, de Lyne Da Sylva, Graham Russel, Yves Marcoux et Frédéric Doll, Équipe du GRDS EBSI Université de Montréal, rend compte de l'expérimentation d'un algorithme de segmentation automatique qui utilise des informations sur la distribution des mots dans le texte pour calculer un score de cohésion lexicale qui peut être utilisé pour repérer des changements de segment thématique. Ce type d'information s'est avéré pertinent pour le corpus juridique. Toujours dans l'optique

d'améliorer les performances des classifieurs, une équipe du LIA Université d'Avignon et des pays du Vaucluse (Eric Charton, Nathalie Camelin, Rodrigo Acuna-Agost, Rémi Lavalley, Rémy Kessler et Silvia Fernandez) met en évidence l'intérêt de combiner des prétraitements par analyse distributionnelle avec une fusion de résultats des meilleurs classifieurs pour la classification automatique de corpus en genres et catégories.

Les campagnes en fouille d'opinions

Cette partie rassemble une contribution de la campagne 2007 « Fouille d'opinions sur des avis argumentés (critiques de livres, de films) », et deux de la campagne 2009, qui comporte trois tâches visant à la distinction objectivité/subjectivité dans un corpus multilingue : 1° Détermination du parti politique d'appartenance des parlementaires d'après leurs interventions, 2° Détection des articles globalement objectifs (factuels) ou subjectifs (porteurs d'opinion) dans les corpus de journaux, et 3° Identification des passages subjectifs (du paragraphe au mot) dans les corpus de débats et de journaux.

La première contribution (Luca Dini, Sigrid Maurel, Paolo Curtoni et Beata Dobrzyńska, Equipe Ho2S (anciennement CELI France)) combine une approche mixte, symbolique et statistique ; la couche symbolique permettant d'améliorer les performances de la couche statistique. Elle donne également un point de vue sur les applications industrielles de cette approche.

Yves Bestgen (FRS-FNRS Belgique) développe une série d'essais d'optimisation d'un classifieur SVM en jouant sur des paramètres couramment utilisés de sélection des descripteurs et d'autres paramètres, moins classiques, tels que la longueur des documents selon la catégorie à laquelle ils ont été affectés.

La troisième contribution (Matthieu Vernier, Laura Monceaux et Béatrice Daille, (LINA Université de Nantes)) est à mon avis la plus aboutie sur le plan théorique et pratique. Elle s'intéresse aux constituants de la subjectivité dans le langage et sur le plan informatique repose sur une architecture UIMA pour l'intégration des composants d'annotation au sein d'une chaîne de traitement.

Les campagnes diachroniques

La campagne 2010 a comme objectifs, d'une part, l'identification de la décennie de publication d'un extrait de journal (de 1800 à 1940), soit 15 décennies, et d'autre part l'identification du pays d'origine (France *vs* Québec) d'un article de journal puis l'identification du journal de l'article étudié. L'un des points problématiques de cette campagne concerne les frontières de classe. En effet un système retournant la décennie 1920 pour un document de 1919 sera pénalisé, alors que l'année de parution est plus proche de la décennie 1920 que de celle de 1910. La campagne 2011 corrige le tir en ayant pour objectif principal l'identification de l'année de publication d'un extrait de journal (de 1801 à 1944). Pour autant, cette modification implique un changement d'échelle en augmentant le nombre de classes (on passe de 15 décennies à 144 années différentes). C'est pourquoi l'évaluation des résultats est basée sur une fenêtre de temps autour de l'année de référence, en évaluant la proximité avec l'année de référence.

Les articles de cette quatrième partie se caractérisent par la variété des méthodes utilisées et combinées (analyse lexicale, analyse probabiliste, apprentissage supervisé, intégration de connaissances sur les réformes orthographiques du français, les archaïsmes et néologismes, etc.). Deux articles me paraissent particulièrement intéressants. Le premier, Apprentissage supervisé et paresseux pour la fouille de textes (Christian Raymond et Vincent Claveau, équipe IRISA-CNRS Insa de Rennes) montre que l'emploi de classifieurs très complexes est contre-performant pour cette tâche spécifique de datation. Le second, Méthodes pour l'archéologie linguistique : datation par combinaison d'indices temporels (Anne García-Fernandez, Anne-Laure Ligozat, Delphine Bernhard et Marco Dinarelli du LIMSI-CNRS France) est intéressant pour son utilisation de ressources externes (Wikipédia, Google Books) pour prendre en compte l'évolution diachronique d'une langue en combinaison avec des techniques classiques de classifications supervisées.

Pour conclure, cet ouvrage satisfera particulièrement l'ingénieur ou l'enseignant-chercheur linguiste, informaticien linguiste, informaticien, spécialiste des techniques d'extraction d'information et/ou de classification supervisées ou non, qui y trouvera à la fois un panorama des méthodes symboliques, statistiques et distributionnelles, une grande variété de techniques de prétraitement des données d'entrée (pour prendre en compte leurs spécificités), mais aussi des arguments pour orienter ses choix dans le développement d'un système à finalité de fouille de textes et enfin une méthodologie pour l'évaluation des performances de ce type de système. Cet ouvrage est également une vitrine de la recherche francophone en ingénierie de la langue et une contribution importante au développement d'un socle technologique et méthodologique francophone pour la fouille de textes. De fortes compétences en linguistique informatique, en classification et analyse de données sont nécessaires pour tirer pleinement profit des richesses des contributions.

Luc Grivel, Maître de Conférences de l'Université de Paris 1 (Panthéon-Sorbonne) et chercheur associé à l'équipe INDEX du Laboratoire Paragraphe de l'Université de Paris 8

Anthony Aycock. *The Accidental Law Librarian.* Medford, NJ: Information Today, 2013. ISBN: 978-1-57387-477-9. CDN$43.00.

Like for many librarians, the idea of legal research sends chills up my spine. Give me an obscure contemporary art question or an overly technical business question any day, but until I finished *The Accidental Law Librarian*, it seemed that legal research and reference assistance was best left to those who had completed a *juris doctor* and possibly even several years as a lawyer or judge. With *The Accidental Law Librarian*, the latest in the *Accidental* series, Anthony Aycock has done an excellent job of making the world of legal resources and research much more accessible to those who do not face these types of questions on a daily basis.